A Network Datagram and Big Data Based Research on Method of User Profile

Jiabin Li [a,*], Zhi Xue [b]

School of Cyber Security, Shanghai Jiao Tong University, 800 Rd. Dongchuan, Shanghai, China

[a] jacksonstarry@sjtu.edu.cn, [b] zxue@sjtu.edu.cn

Keywords: Big Data, User Profile, Network Datagram, Behaviour Forecast, Security Management.

Abstract. Big data analysis technology is getting more and more popular among various enterprises, government departments and other institutions. With the help of huge amount of data, network users' behaviour in cyber space can be vividly described. This paper first introduces the status of research on user profile, then puts forward a new method, using big data technology to analysis users' network datagram; then the paper builds up a system to wrap the data storage and search module, the profiling module and visualizing module all together. The paper aims to provide some insights for better governance, enterprise operation and security management.

1. Introduction

Recent years, big data analysis has become a significant and popular technology among more and more enterprises, Government departments and organizations. By analyzing huge amount of network traffic datagrams, a network user's internet surfing profile can be figured, including habits, preferences, consumption level, active online periods, etc. Furthermore, accumulating the datagram in the dimension of both time and cyber topological space can build behaviour model to forecast online consumers' or network users' future behaviour as well as to design a better optimized software or applications.

Not only does big data analysis provide a unique tool in livelihood information collection[1] and consumer behaviour forecast, but also it works well in cyber security area. Security departments can monitor and analyze users' daily log of internet surfing to get a real-time alert of abnormal network traffic recognition. In the meanwhile, enterprises can also derive valid user behaviour model by analyzing massive users network data, thus to intercept or filter malicious access requests in advance.

Another usage of big data analysis is to form a long-term tracking analysis to the target user or network node, and then generate a complete individual behaviour model, and group one, by comparing plural individual models. Thus, the main concern is to build up an effective user behaviour analysis system based on big data platform.

2. Method of Analyzing User Behaviour Data

User behaviour can be profiled mainly by analyzing, deconstructing and reconstructing the data from raw network datagram[2]. Some common segments are given in Table 1.

Table 1. Common segments in TCP packet.

Segment (Protocol Type)	Meaning
Method (HTTP)	Actions that the client want server to act at the resources, always together with resource path and protocol version
Host (HTTP)	The host and port number of network resource requested.
User-Agent (HTTP)	Containing visitor's device information, browser version and OS version, etc.
Content-Type (HTTP)	Type of resource responded
Src/Dst (TCP)	Source & destination IP address
Src/Dst Port (TCP)	Source & destination port number
Src/Dst (Ethernet II)	Source & destination MAC address
Content (HTTP)	Content text of an HTTP packet
Title (HTML Content Text)	Web page title
Keywords (HTML Content Text)	Searching keywords

According to Table 1, we can easily describe when and where a packet is sent, be aware of the data requirement and responding, and even get further analysis result of the content text of a HTTP packet. For example, the "title" tag in HTTP content tells the title of a webpage visited, so as to make analysis more detailed[3]. Thus, a relation model can be concluded as following:

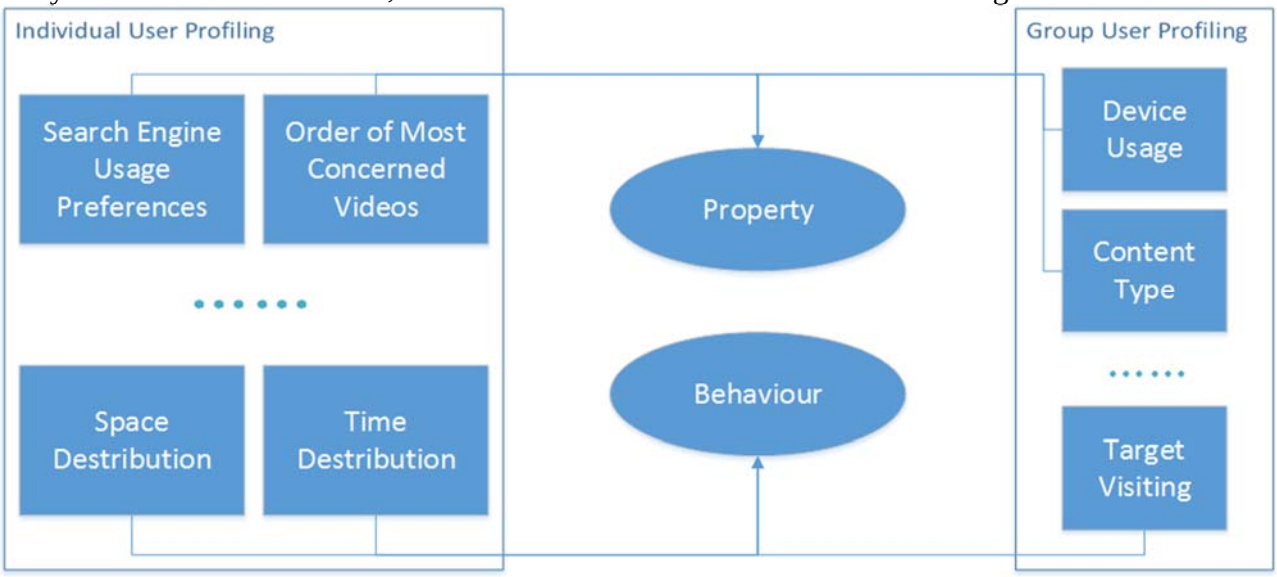

Figure 1. Individual/Group user profile model.

The User profiling could be divided into two sorts as *property* and *behaviour*. Property describes static status, such as consumption level, most concerned topics; Behaviour tells what or how a user or user group tends to do[4].

The former helps figure user's general image, so that users can be sorted or clustered into different groups; the latter dynamically shows user's term characteristics and helps find the behaviour patterns and trends. To get an all-round image of user, the two models should be both considered in specific analysis. The following is the analysis method based on Table 1.

2.1 Examples of Overall Network Platform Oriented Analysis Method.

Overall network platform oriented analysis can profile the behaviour track of one group or more, and get its/their behaviour mode by weighing the users' data.

2.1.1 Distribution of Visited Hostname.

Objective: To list the most welcomed websites by sorting the visited hostname in the platform.

Strategy: Sort the aggregation result of Host segment in descending order.

2.1.2 Mobile Device Usage.

Objective: To identify and classify the most popular mobile device.

Strategy: Filter the User-Agent segment, then select the top x results and sort in descending order.

2.1.3 Webpage Content.

Objective: To select the most concerned topics by analyzing the webpage contents within the whole network platform.

Strategy: First do aggregation on the Title segment, sort the result as the hot topic tags and filter all nonsense results such as “404 Not Found”,”302 Found”, etc.

2.2 Examples of Single Target Oriented Analysis Method.

Other than the overall platform, research can also focus on a single user by analyzing the source/destination IP, MAC and port number and other segments, so that a user’s behaviour pattern can be described completely.

There are more things can be done if comparing the behaviour pattern with existing attack models. For example, if someone posts specific network requests to a certain server abnormally or maliciously in a continuous period, then security manager should mark that user as a suspected attacker, and check whether a network attack really exists.

2.2.1 Time Distribution of Single Source IP.

Objective: To analyze all the request packets from a source IP address and check whether the dataflow is abnormal in the degree of time distribution. For example, if the network traffic generated in a certain Monday is far huger than the general Monday traffic, then take it as an event.

Strategy: Aggregate timestamp segment with date histogram method, and set the set time scale frequency as weekly. Then conduct a sub-aggregation on timestamp segment again, and scale frequency set as hourly. At last wrap a filter to specify the user to be analyzed, such as taking the “src_ip” segment as filter segment, then the time distribution results of all network packets derived from the certain source IP address would be concluded.

2.2.2 Space Distribution of Single Source IP.

Objective: To analyze all the request packets from a source IP address and check whether the dataflow is abnormal in the degree of space distribution. For example, if the destination IP or hostname always drops in some specific clusters, but some day the user conducts a lot of requests to strange servers, then take it as an event.

Strategy: Aggregate timestamp segment by date histogram method, and set the time scale as weekly. Then conduct a sub-aggregation on “dst_ip” or “host” segment, and limit the result size. At last wrap a filter to specify the user to be analyzed, such as taking the “src_ip” segment as filter segment, then the space distribution results of all network packets derived from the certain source IP address would be concluded.

2.2.3 Search Engine Usage Preferences.

Objective: To identify what search engine a user uses most.

Strategy: Filter the host segments with a prepared search engine host list, filter the specific user’s source address.

2.2.4 Most Concerned Videos.

Objective: To collect the users’ favourite video platform and describe the users’ browsing habits

Strategy: First aggregate the Title segment to get a result set of website titles, then nest a filter of Host segment to limit the result to the video providers, then use filter again to specify the user.

3. Big Data Based User Profile Analysis Method and System Design

3.1 Big Data Based User Profile Analysis Method

What Big Data Based User Profile Analysis Method[5] says is to collect, filter and analyze massive data conducted by all users (or network nodes) under the support of big data technology, and finally carry out the relative between user data and user behaviour, thus generate a user profile model or a behaviour trend forecast report.

The user profile model contains users’ online habits, preferences, consumption level, normal routine, etc. While the behavior trend focuses on areal data flow trend, website or application usage growth, incident count fluctuation, etc.

3.2 Big Data Based User Profile Analysis System Design

This paper designs a big data based user profile analysis system that can store and process massive data to analyze all network data generated by all users according to the realistic demands and application environment, then display the data on a real-time visualization website[6].

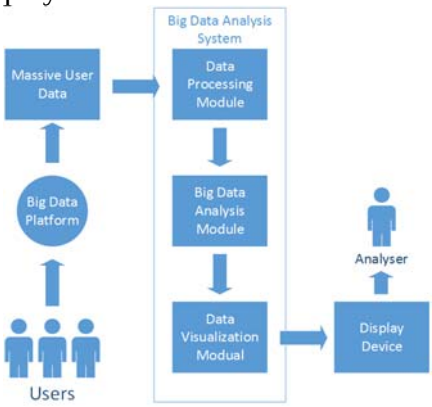

Figure 2. Big Data Analysis Stages and System.

The capture modules capture and pre-process data flow at every network node and send the pre-processed data to the Hadoop platform, stored in HBase. The analysis module starts at each time interval the analyzer set, and dumps the HBase data into JSON format, and imports the JSON data to an open-source distributed document search engine and display on a website.

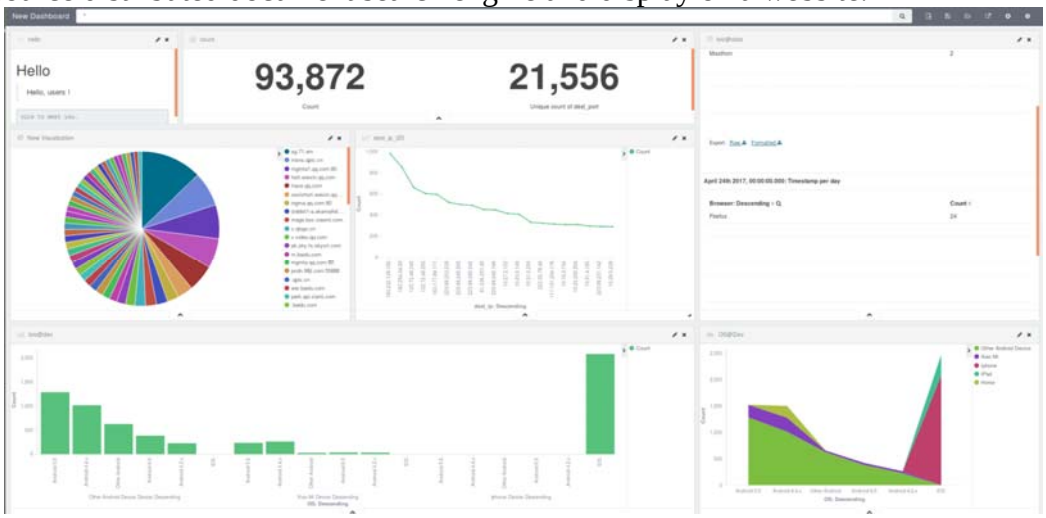

Figure 3. Data Visualization Frame Example.

4. Test and Analysis of the System

4.1 Test: Overall Network Platform Oriented Analysis

4.1.1 Distribution of Visited Hostname

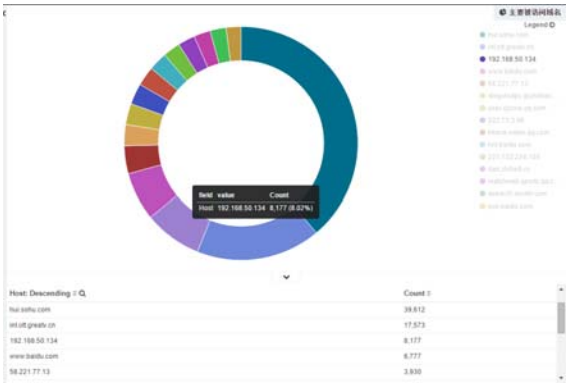

Figure 4. Distribution of Visited Hostname.

According to Figure 4, the most visited host is “soho.com” with the proportion of nearly 40%.

4.1.2 Mobile Device Usage

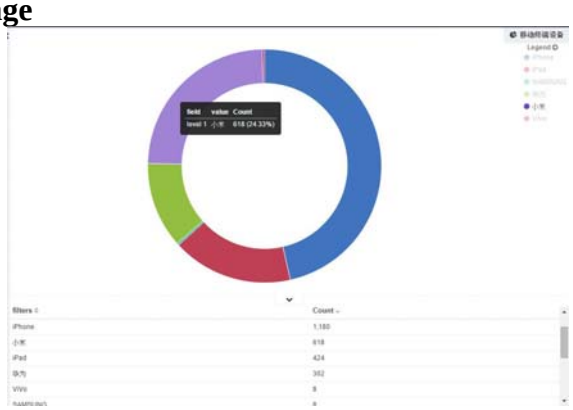

Figure 5. Mobile Device Usage.

According to Figure 5, the top used mobile device is “iPhone” with the quantity of 1,180; the second is “Xiaomi” with the proportion of 24.33%.

4.1.3 Webpage Content

网页访问内容

Title: Descending	Count
Object moved	334
客户订单产品	164
visual addImageQuery="true" baseUri="http://cn.img.s-msn.com/tenant/amp/entityid/">	124
订单出库	88
visual version="2" lang="zh-CN">	76
New Document	72
(盘中特供)厦门钨业动力锂电池材料生产线将扩大3倍至18条	72
Apache Tomcat/5.5.27 - Error report	66
打印订单信息	62
SU_TITLE	52

Export: Raw Formatted

1 2 3 4 5 »

Figure 6. Webpage Content.

According to Figure 6, the most visited meaningful webpage is about “Customer Order” from an e-commerce site, and the next is “Order Delivery”.

4.2 Test: Single Target Oriented Analysis Method Examples

4.2.1 Time Distribution of Single Source IP.

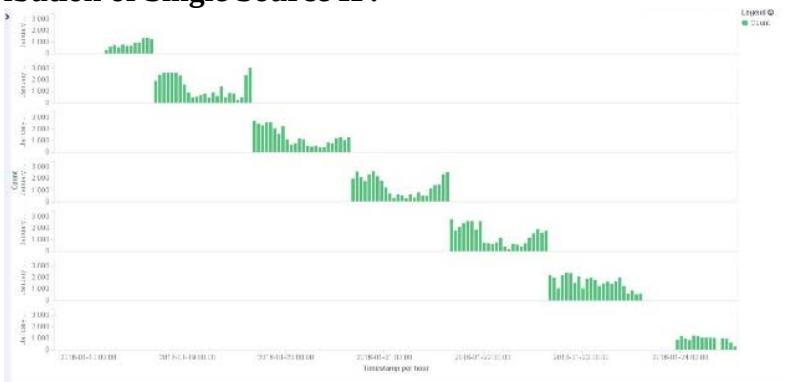

Figure 7. Time Distribution of Single Source IP.

According to Figure 7, the user whose IP is “10.161.35.249” conduct an active online period from 11 p.m. till the next 6 a.m., while a nearly zero data generation from 9 a.m. till 5 p.m. And during weekends, the quantity of dataflow keeps high before 10 p.m., which means the user is possibly running some software or applications that keep connecting to the internet background, and his active period is afternoon.

4.2.2 Space Distribution of Single Source IP.

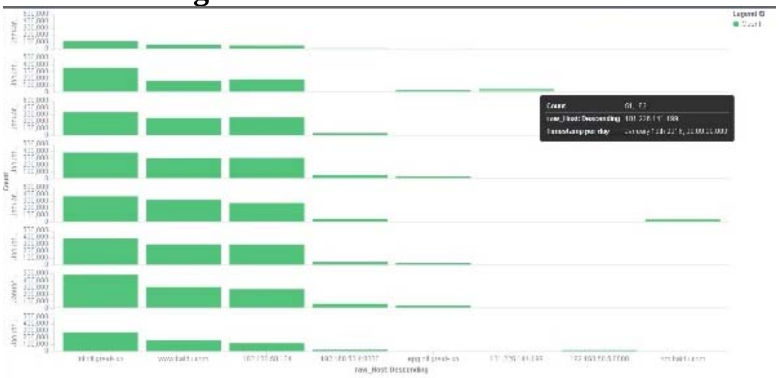

Figure 8. Space Distribution of Single Source IP.

According to Figure 8, the user whose MAC is XXXX(hide for personal privacy), visits “int.ott.gratv.cn”,”www.baidu.com”,”192.168.50.134” most, but on 18th Jan. huge amount of visit to 101.226.141.199 occurred, which should arouse security manager’s alert.

4.2.3 Search Engine Usage Preferences.

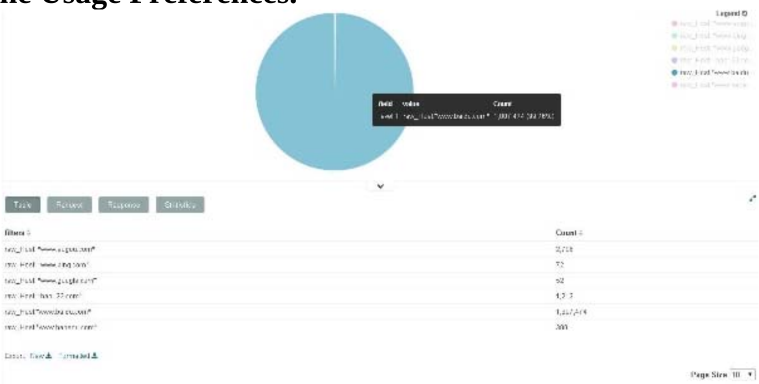

Figure 9. Search Engine Usage Preferences.

According to Figure 9, the user whose IP is 10.226.141.199 uses “baidu” as his or her search engine most, with the proportion of99.76%; the next is “sogou”, with the proportion of0.15%.

4.2.4 Most Concerned Videos.

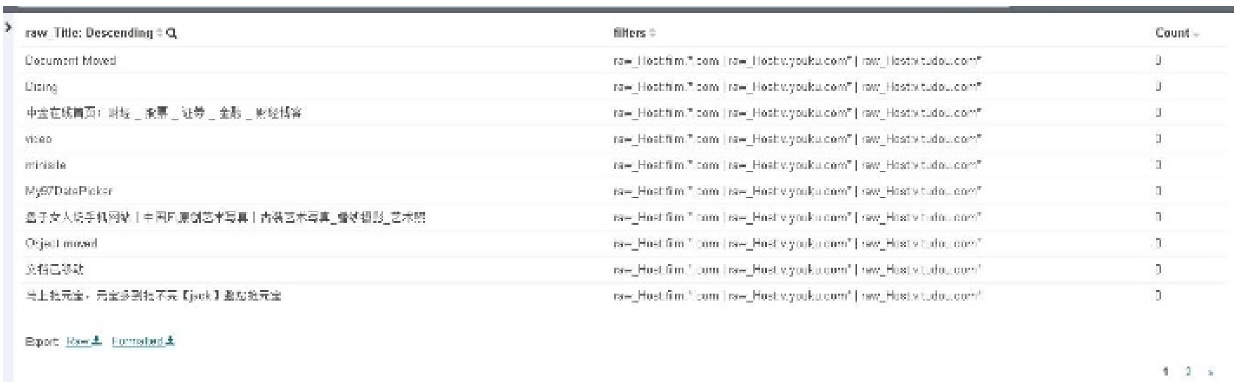

Figure 10. Most Concerned Videos.

According to Figure 10, the user whose IP is 10.161.35.249 watches “ China Economy Online” most.

5. Conclusion

Indeed, big data analysis can be divided into two types: one is to conclude the already occurred data into some regulations, the other is to predict, to forecast the future trend. Due to the 4V characteristic, big data has been generally accepted to be the key technology in future analysis area[7].

This paper introduces an analysis method based on datagram and big data technology together with a visualization system, so as to conduct a user (group) behaviour profile and behaviour trend, which is designed to support better governance, enterprise operation and security management.

The result this paper concludes can be adapted in many realistic application scenes. The analysis and statistic resulting from TCP header and non-https contents are accurate, but the data, encrypted or sent via covert channels, are still hard to analyze, and which might be the main research area of author in the next phase of research.

Acknowledgements

I would like to express my gratitude to all those who have helped me during the writing of this paper. I gratefully acknowledge the help of my supervisor Professor Xue Zhi. I do appreciate his patience and professional instructions during my paper writing. Also, I would like to thank my wife Shao Yue, who gave me encouragement and helped to proofread this paper.

Last but not the least, my gratitude also extends to my family who have been assisting, supporting and caring for me all of my life.

This work was supported by the Key Program of the Natural Science Foundation of China (No. 61332010)

References

[1]. Zhang Lu, Li Xiaoyong, Ma Wei. (2014) Research on the Security Model of Big Data in Government. Netinfo Security, 5, 63-67.
[2]. Jiang Kaida, Li Xiao, Sun Qiang. (2014) Big Data Analysis on Security based on Network Traffic Metadata. Netinfo Security, 5, 37-40.
[3]. Chen Cheng. (2015) Research on Personalized Library Services and User Behaviour Analysis Based on Big Data. Library Work and Study, 2, 28-31.
[4]. Hu Yuchen, Guo Yu. (2013) Mobile Network User Behaviour Analysis Based on Hourglass Model. Management World, 7, 84-185.
[5]. Ren Siying. (2014) Network User Behaviour Analysis Based on Big Data. Beijing University of Posts and Telecommunications, 12.
[6]. Tao Caixia, XieXiaojun, Chen Kang, Guo Lirong, Liu Chun. (2013) Design of Mobile Internet Big Data User Behavior Analysis Engine Based on Cloud Computing. Telecommunications Science, 3, 27-31.
[7]. Chen Jianchang. (2013) Analysis of Network Security in Big Data Environment. China New Telecommunications, 17, 13-16.